ASA Statement on Statistical Significance and *P*-Values

1. Introduction

Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions drawn from research data. The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.

Underpinning many published scientific conclusions is the concept of "statistical significance," typically assessed with an index called the *p*-value. While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of *p*-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since *p*-values were first introduced.

In this context, the American Statistical Association (ASA) believes that the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value. The issues touched on here affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law. This statement does not seek to resolve all the issues relating to sound statistical practice, nor to settle foundational controversies. Rather, the statement articulates in nontechnical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community.

2. What is a *p*-Value?

Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

3. Principles

1. ***P*-values can indicate how incompatible the data are with a specified statistical model.**

   A *p*-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the *p*-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2. ***P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**

   Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

3. **Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.**

   Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that *p*-values alone can ensure that a decision is correct or incorrect. The widespread use of "statistical significance" (generally interpreted as "$p \leq 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

4. **Proper inference requires full reporting and transparency**

   P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference, and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. One need not formally carry out multiple statistical tests for this problem to arise: Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed. Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting.

5. **A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.**

   Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs.

6. **By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.**

   Researchers should recognize that a p-value without context or other evidence provides limited information. For example, a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large p-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible.

4. Other Approaches

In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates. All these measures and approaches rely on further assumptions, but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct.

5. Conclusion

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

Acknowledgments

The ASA Board of Directors thanks the following people for sharing their expertise and perspectives during the development of the statement. The statement does not necessarily reflect the viewpoint of all these people, and in fact some have views that are in opposition to all or part of the statement. Nonetheless, we are deeply grateful for their contributions. Naomi Altman, Jim Berger, Yoav Benjamini, Don Berry, Brad Carlin, John Carlin, George Cobb, Marie Davidian, Steve Fienberg, Andrew Gelman, Steve Goodman, Sander Greenland, Guido Imbens, John Ioannidis, Valen Johnson, Michael Lavine, Michael Lew, Rod Little, Deborah Mayo, Chuck McCulloch, Michele Millar, Sally Morton, Regina Nuzzo, Hilary Parker, Kenneth Rothman, Don Rubin, Stephen Senn, Uri Simonsohn, Dalene Stangl, Philip Stark, Steve Ziliak.

Edited by Ronald L.Wasserstein, Executive Director
On behalf of the American Statistical Association Board of Directors

A Brief *p*-Values and Statistical Significance Reference List

Altman D.G., and Bland J.M. (1995), "Absence of Evidence is not Evidence of Absence," *British Medical Journal*, 311, 485.

Altman, D.G., Machin, D., Bryant, T.N., and Gardner, M.J. (eds.) (2000), *Statistics with Confidence* (2nd ed.), London: BMJ Books.

Berger, J.O., and Delampady, M. (1987), "Testing Precise Hypotheses," *Statistical Science*, 2, 317–335.

Berry, D. (2012), "Multiplicities in Cancer Research: Ubiquitous and Necessary Evils," *Journal of the National Cancer Institute*, 104, 1124–1132.

Christensen, R. (2005), "Testing Fisher, Neyman, Pearson, and Bayes," *The American Statistician*, 59, 121–126.

Cox, D.R. (1982), "Statistical Significance Tests," *British Journal of Clinical Pharmacology*, 14, 325–331.

Edwards, W., Lindman, H., and Savage, L.J. (1963), "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science [online]," *American Scientist*, 102. Available at *http://www.americanscientist.org/issues/feature/2014/6/the-statistical-crisis-in-science*

Gelman, A., and Stern, H.S. (2006), "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant," *The American Statistician*, 60, 328–331.

Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socioeconomics*, 33, 567–606.

Goodman, S.N. (1999a), "Toward Evidence-Based Medical Statistics 1: The *P*-Value Fallacy," *Annals of Internal Medicine*, 130, 995–1004.

——— (1999b), "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor," *Annals of Internal Medicine*, 130, 1005–1013.

——— (2008), "A Dirty Dozen: Twelve *P*-Value Misconceptions," *Seminars in Hematology*, 45, 135–140.

Greenland, S. (2011), "Null Misinterpretation in Statistical Testing and its Impact on Health Risk Assessment," *Preventive Medicine*, 53, 225–228.

——— (2012), "Nonsignificance Plus High Power Does Not Imply Support for the Null Over the Alternative," *Annals of Epidemiology*, 22, 364–368.

Greenland, S., and Poole, C. (2011), "Problems in Common Interpretations of Statistics in Scientific Articles, Expert Reports, and Testimony," *Jurimetrics*, 51, 113–129.

Hoenig, J.M., and Heisey, D.M. (2001), "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," *The American Statistician*, 55, 19–24.

Ioannidis, J.P. (2005), "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," *Journal of the American Medical Association*, 294, 218–228.

——— (2008), "Why Most Discovered True Associations are Inflated" (with discussion), *Epidemiology* 19, 640–658.

Johnson, V.E. (2013), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Scienc*es, 110(48), 19313–19317.

——— (2013), "Uniformly Most Powerful Bayesian Tests," *Annals of Statistics*, 41, 1716–1741.

Lang, J., Rothman K.J., and Cann, C.I. (1998), "That Confounded *P*-value" (editorial), *Epidemiology*, 9, 7–8.

Lavine, M. (1999), "What is Bayesian Statistics and Why Everything Else is Wrong," *UMAP Journal*, 20, 2.

Lew, M.J. (2012), "Bad Statistical Practice in Pharmacology (and Other Basic Biomedical Disciplines): You Probably Don't Know *P*," *British Journal of Pharmacology*, 166, 5, 1559–1567.

Phillips, C.V. (2004), "Publication Bias In Situ," *BMC Medical Research Methodology*, 4, 20.

Poole, C. (1987), "Beyond the Confidence Interval," *American Journal of Public Health*, 77, 195–199.

——— (2001), "Low *P*-values or Narrow Confidence Intervals: Which are More Durable?" *Epidemiology*, 12, 291–294.

Rothman, K.J. (1978), "A Show of Confidence" (editorial), *New England Journal of Medicine*, 299, 1362–1363.

——— (1986), "Significance Questing" (editorial), *Annals of Internal Medicine*, 105, 445–447.

———- (2010), "Curbing Type I and Type II Errors," *European Journal of Epidemiology*, 25, 223–224.

Rothman, K.J., Weiss, N.S., Robins, J., Neutra, R., and Stellman, S. (1992), "Amicus Curiae Brief for the U. S. Supreme Court, *Daubert v. Merrell Dow Pharmaceuticals*, Petition for Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit," No. 92-102, October Term, 1992.

Rozeboom, W.M. (1960), "The Fallacy of the Null-Hypothesis Significance Test," *Psychological Bulletin*, 57, 416–428.

Schervish, M.J. (1996), "*P*-Values: What They Are and What They Are Not," *The American Statistician*, 50, 203–206.

Simmons, J.P., Nelson, L.D., and Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22, 1359–1366.

Stang, A., and Rothman, K.J. (2011), "That Confounded *P*-value Revisited," *Journal of Clinical Epidemiology*, 64, 1047–1048.

Stang, A., Poole, C., and Kuss, O. (2010), "The Ongoing Tyranny of Statistical Significance Testing in Biomedical Research," *European Journal of Epidemiology*, 25, 225–230.

Sterne, J. A. C. (2002). "Teaching Hypothesis Tests—Time for Significant Change?" *Statistics in Medicine*, 21, 985–994.

Sterne, J. A. C., and Smith, G. D. (2001), "Sifting the Evidence—What's Wrong with Significance Tests?" *British Medical Journal*, 322, 226–231.

Ziliak, S.T. (2010), "The Validus Medicus and a New Gold Standard," *The Lancet*, 376, 9738, 324–325.

Ziliak, S.T., and McCloskey, D.N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press.